Selection/Extraction of Spectral Regions for Autofluorescence Spectra Measured in the Oral Cavity

Marina Skurichina[1], Pavel Paclík[1], Robert P.W. Duin[1], Diana de Veld[2], Henricus J.C.M. Sterenborg[2], Max J.H. Witjes[3], and Jan L.N. Roodenburg[3]

[1] Information and Communication Theory Group, Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, P.O. Box 5031, 2600GA Delft, The Netherlands
m.skurichina@ewi.tudelft.nl
[2] Photodynamic Therapy and Optical Spectroscopy Programme, Department of Radiation Oncology, Erasmus MC, Rotterdam, The Netherlands
[3] Department of Oral and Maxillofacial Surgery, University Hospital Groningen, The Netherlands

Abstract. Recently a number of successful algorithms to select/extract discriminative spectral regions was introduced. These methods may be more beneficial than the standard feature selection/extraction methods for spectral classification. In this paper, on the example of autofluorescence spectra measured in the oral cavity, we intend to get deeper understanding what might be the best way to select informative spectral regions and what factors may influence the success of this approach.

1 Introduction

In medical applications, one often faces the small sample size problem: the number of measurements is smaller than or comparable with the data dimensionality. In such conditions, it is difficult (or even impossible) to construct a good classification rule [1]. One has to reduce the data dimensionality. When having spectral data, using standard feature selection/extraction methods may be inconvenient. The standard approaches assume the independency of data features while in spectra the features (neighbouring wavelengths/pixels/bins) are correlated. Therefore, some useful information may be lost if the connectivity of spectral neighbouring pixels is not taken into account when extracting/selecting features informative for discrimination between data classes. During the last few years, a number of novel methods for selection/ extraction informative spectral regions/bands has been developed. One example of such feature extraction algorithms is an Optimal Region Selector (ORS) [2] guided by a genetic algorithm. Another example is a top-down multiresolution feature extraction algorithm proposed by Kumar, Ghosh and Crawford [7].

In this paper we pretend neither to introduce fundamentally new algorithms for selection/extraction of informative spectral regions, nor to perform an extensive comparison of already existing algorithms with a standard feature selection/extraction methods. Our goal is to understand what happens exactly when extracting informative spectral regions by different methods, what underlines the success of this process and what factors influence the benefit of this approach.

A. Fred et al. (Eds.): SSPR&SPR 2004, LNCS 3138, pp. 1096–1104, 2004.

In order to perform our study we have selected a real data set, which represents autofluorescence spectra measured in the oral cavity. This data set introduces a 2-class problem: lesions against healthy tissues. It is described in section 2. Different feature selection/extraction techniques used for finding informative spectral regions/bands are introduced in section 3. The results of our simulation study are presented in section 4. Conclusions can be found in section 5.

2 Data

We perform our study on the example of autofluorescence spectra measured in the oral cavity. The data consist of the autofluorescence spectra acquired from healthy and diseased mucosa in the oral cavity. The measurements were performed at the Department of Oral and Maxillofacial Surgery of the University Hospital of Groningen [3]. Autofluorescence spectra were collected from 97 volunteers with no clinically observable lesions of the oral mucosa and 137 patients having lesions in the oral cavity. The measurements were taken at 11 different anatomical locations with excitation wavelength equal to 365 nm. The previously performed study [3] has shown that spectra measured at different anatomical locations are similar, and the location of a probe affects only the intensity of the spectra but not the shape. By this, it was possible to use a larger data set for our study. In total, 856 spectra representing healthy tissue and 132 spectra representing diseased tissue were obtained. After preprocessing [3], each spectrum consists of 199 bins (pixels/wavelengths).

In order to get rid of a large deviation in a spectral intensity within each data class, we normalized spectra by the Unit Area (UA)

$$a_i^{UA} = \frac{a_i}{U}, \quad U = \sum_{j=1}^{199} a_j, \quad i = 1, \ldots, 199, \tag{1}$$

where a_i is an intensity of a spectrum $\boldsymbol{A}=\{a_1, ..., a_{199}\}$ at bin i, i=1,...,199. Normalized autofluorescence spectra representing healthy and diseased tissues and their median spectra are illustrated in Fig. 1.

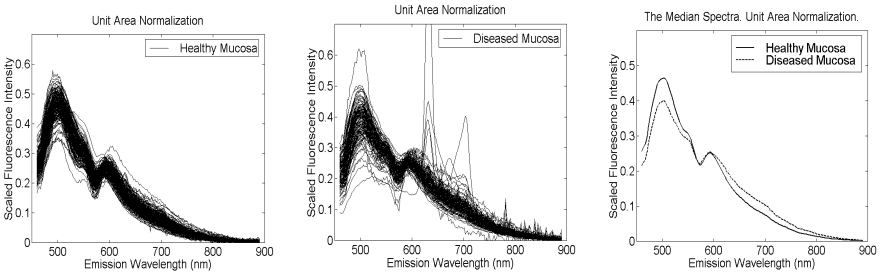

Fig. 1. Normalized autofluorescence spectra for healthy and diseased mucosa in oral cavity.

For our simulation study, training data sets with 2/3 of available samples per class are chosen randomly from the total set. The remaining data are used for testing. The prior class probabilities are set to be equal as the data are very unbalanced and the real prior class probabilities are unknown. To evaluate the performance of lesion diagnos-

tics when different feature selection/extraction methods are used, we have chosen the Linear Discriminant Analysis (LDA) [4] which was the best performing classifier for this application. In particular, we apply the regularized linear classifier [5] which constructs a linear discriminant function assuming normal class distributions and using a joint class covariance matrix for both data classes. The value of the regularization parameter used is equal to 10^{-10}. All experiments are repeated 20 times on independent training sample sets. In all figures the averaged results over 20 trials are presented and we do not mention that anymore. The standard deviations of the reported mean generalization errors (the mean per two data classes) is approximately 0.01 for each considered case.

3 Selection/Extraction of Spectral Regions

Inspired by the success of approaches suggested by Nikulin [2] and Kumar [7], we became interested in what actually happens when selecting/extracting spectral regions, why and when it is beneficial and not.

The first approach, Optimal Region Selector by Nikulin [2], is based on a genetic algorithm. First, one randomly generates few sets of non-overlapping spectral regions of arbitrary size. For each region, a new feature (for instance, the mean of the spectral intensities in the region) is derived. Then the goodness of each set of new features is evaluated by some criterion (Nikulin has used the mean square error between the true labels and the posterior class probabilities calculated on the training dataset by the linear classifier having the averaged coefficients over all linear classifiers constructed on leave-one-out cross-validation). According to this criterion, the best subsets of spectral regions are selected. Further, one again increases the number of these subsets by random mutations and crossovers of the region definitions, and the procedure repeats. At the end, a suboptimal solution (due to a randomness of the whole procedure) for the best set of spectral regions is found.

The second approach, a top-down multiresolution feature extraction algorithm proposed by Kumar et al. [7], partitions the original p-dimensional spectra into smaller subspaces by using a top-down recursive algorithm. First, the best place to split spectra into two parts is found by computing a discriminant measure between data classes (for instance, Bhattacharya distance, Kullback-Leibler divergence or log-odds of class posterior probabilities used by Kumar can be applied). The discriminant measure obtained on the parent space is compared with the discriminant measures calculated on the children subspaces. If the child subspace has a higher discrimination than the parent space, then it is partitioned further. If the child subspace does not show any improvement in its discrimination capacity compared to the parent space, then this child subspace is not partitioned any further. Finally, one finds a set of spectral regions/ bands with high discrimination. However, the optimization is performed only in a one-dimensional way: a discrimination capacity is evaluated for each spectral region separately but not for a total set of selected spectral regions.

We consider here the Top-Down variant of Generalized Local Discriminant Bases algorithm (GLDB-TD), which is conceptually close to the algorithms, described in this paper. The GLDB-TD algorithm represents each group of wavelengths by a mean of corresponding intensities.

In our study on the usefulness of spectral regions extraction/selection approach and on the benefits of different ways to extract/select discriminative spectral regions, we

do not use exactly the two algorithms discussed above. First, we would like to simplify the procedures for selection/extraction of discriminative spectral regions in order to get more insight what does actually happen. Second, it is convenient to use the same discriminant measure between data classes when comparing different ways of selecting/extracting the best spectral bands. And finally, we prefer to use multi-dimensional optimization when looking for the informative spectral regions.

In our study we consider few approaches to extract/select the informative/discriminative spectral regions. In all of them, first we perform the dimensionality reduction for each considered spectral band. Namely, from each spectral region considered we derive one new feature by taking the average of intensities in this region. This new feature (the averaged intensity of the spectral band) is used further to introduce the spectral region.

In order to evaluate a discriminative capacity of extracted spectral regions, we use the Mahalanobis Distance (MD) between data classes:

$$MD = (\mu_A - \mu_B)'(p\Sigma_A + (1-p)\Sigma_B)^{-1}(\mu_A - \mu_B), \quad (2)$$

where μ_A, μ_B and Σ_A, Σ_B are the means and the covariance matrices of data classes A and B, respectively; p is the prior probability of the data class A. The larger Mahalanobis distance, the larger discriminative capacity between data classes. In order to perform the multi-dimensional optimization of S spectral regions, we calculate the Mahalanobis distance on the whole set of S features (the averaged intensities of spectral bands), each representing one of S spectral regions. By this, we find the optimal set of spectral regions providing the best discrimination (according to MD) between data classes.

In our study we consider the following ways to extract/select the informative spectral regions.

Approach 1.

A. *Sequential Partition of Spectra into Non-overlapping Bands Using All Spectral Pixels.*

First, we split spectra into two spectral regions by finding the best split which gives the largest MD (over all possible partitions) in the space of two features extracted from the two spectral bands. Then, the first found split is fixed and we look for the next optimal split in such a way that the MD in a three-dimensional space (on three features extracted from the three spectral bands) is the largest over all possible locations for the second split (when the location of the first split is anchored). Again, we fix the location for the first two found splits and repeat the procedure while the desired number of spectral regions S is found (see top plots in Fig. 2). In this approach, all spectral emission wavelengths are used in the partitioning of spectra. However, some spectral bins can be uninformative - introducing only noise. Hence, they may deteriorate the classification when they are included in the extracted spectral regions. Therefore, it is good to remove them from the spectral bands. One way to do this is described below.

B. *Sequential Partition of Spectra into Non-overlapping Bands Excluding Uninformative Spectral Pixels.*

After a desired number of spectral regions S is found by Approach 1A, we can shrink the spectral bands removing uninformative emission wavelengths. We reduce the

number of bins in each spectral region in a sequential way moving from the most left spectral region to the most right one. For shrinking the spectral band, we consider all possible subregions of the reduced size in this band and find the one with the largest MD in S-dimensional space (one feature, the averaged intensity, calculated from a shrunk subregion of the spectral band under consideration and the rest S-1 features extracted from the other S-1 spectral bands which definitions are fixed for a moment). After the optimal shrinking for the first spectral band is found, we anchor its new definition and move to the next spectral band in order to exclude uninformative pixels (see middle plots in Fig. 2). We should mention that the proposed method is highly dependent on the spectral regions proceeding order and therefore it does not guaranty the optimal shrinking for all regions in general.

Approach 2.

A. *Sequential Selection/Extraction of Discriminative Spectral Regions.*

In order to find a set of the most discriminative spectral bands, at each step s, s = 1, 2, … S, we consider all possible definitions of spectral regions (of arbitrary size) in spectra. For each of them we calculate the MD criterion in s dimensions: one

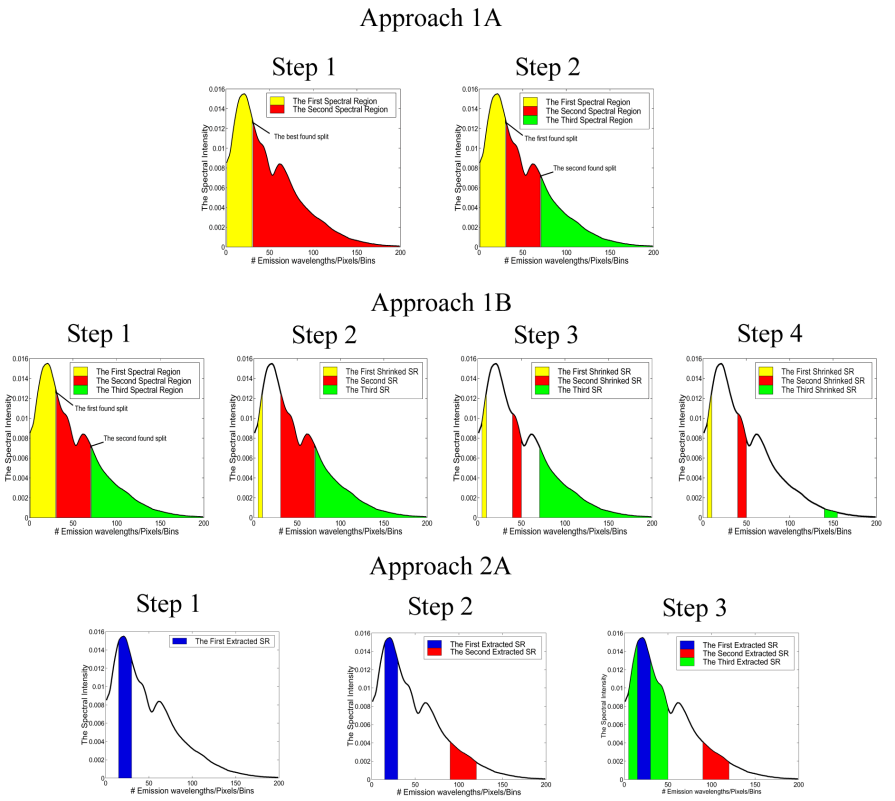

Fig. 2. Approach 1A (top plots), Approach 1B (middle plots) and Approach 2A (bottom plots) for selection/extraction of informative spectral regions (SR) to discriminate between healthy and diseased tissues.

feature is the averaged intensity of a current potential pretender for the most informative band and other s-1 features are extracted from the previously found optimal spectral regions. The spectral band (a potential pretender) with the largest MD is picked as the most discriminative spectral band (in combination with the s-1 previously found optimal regions). Let us mention that in this approach overlapping as well as non-overlapping spectral bands are possible (see bottom plots in Fig. 2).

B. *Sequential Selection/Extraction of Non-overlapping Discriminative Spectral Regions.*

This approach is identical to Approach 2A with the exception that overlapping spectral bands are not allowed: when looking for an additional discriminative spectral region, the regions overlapping with the previously selected optimal spectral bands are excluded from consideration.

4 Simulation Study

Let us now study the benefits of extracting/selecting the discriminative spectral regions. In order to judge the success of this approach, we compare different ways to extract/select spectral regions with one of the most successful standard feature extraction methods - the Principal Component Analysis (PCA) [4]. In Fig. 3, we present the mean generalization errors of the LDA (top plots) (over 20 independent trials) and the mean Mahalanobis distances (bottom plots) obtained on the autofluorescence spectral

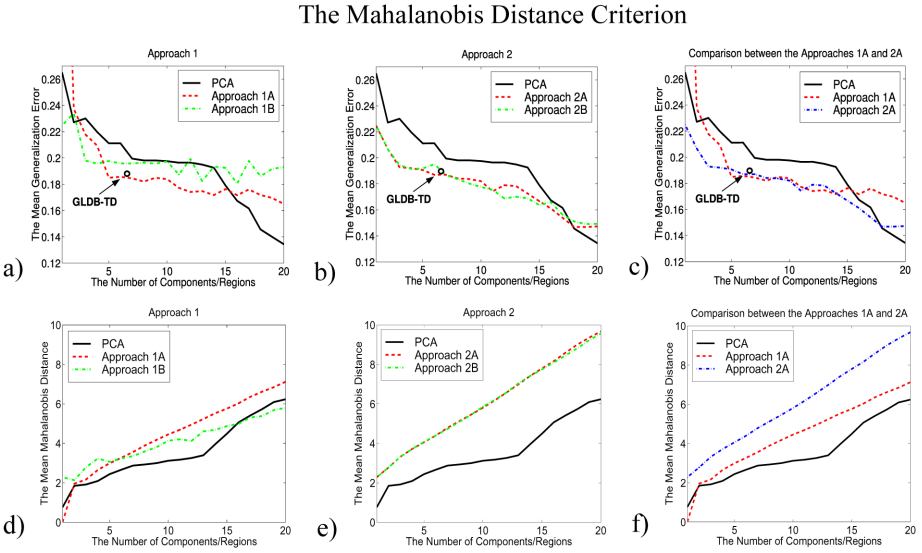

Fig. 3. The mean generalization error (GE) of LDA (a-c) and the mean Mahalanobis distance (MD) (d-f) when the MD criterion is used in Approaches 1 (a,d) and 2 (b,e) to select/extract discriminative spectral bands for autofluorescence spectral data. In plots (c,f) the comparison is made between Approaches 1A and 2A. The standard deviation of the mean GE is around 0.01. GLDB-TD denotes the performance of a Top-Down variant of the Generalized Local Discriminant Bases using a Mahalanobis distance criterion.

data when feature extraction is performed by the PCA and Approaches 1A, 1B, 2A and 2B for discriminative spectral bands extraction. We see that almost all introduced approaches for selection/extraction of informative spectral regions outperform the PCA. The exceptions are the cases when the whole spectrum is taken as one spectral band (the averaged intensity over the whole spectrum is not very informative) or too many spectral regions (some of them contain only noise) are picked to discriminate between data classes.

In order to evaluate our approach we compare different feature extraction techniques, proposed by us, to two existing methods. The first is a Principal Component Analysis (PCA). The second is a Top-Down Generalized Local Discriminant Bases algorithm (GLDB-TD) of Kumar et al. [7].

Similarly to the proposed feature extraction techniques, the GLDB-TD algorithm uses the Mahalanobis distance as a criterion.

Because the GLDB-TD algorithm terminates automatically using a data-driven criterion, only a single point is given in each plot. The point represents the mean error of 20-fold cross-validation. Because each fold results, in general, in a different number of wavelength groups, we plot a median over these 20 results.

Comparing the performance of the linear classifier for Approaches 1A and 1B, we notice that shrinking the spectral regions in the optimal partition of spectra is not useful (see Fig. 3a). One reason underlies in the proposed method 1B that is not optimal in general. Another reason is that the MD criterion is not equivalent to the LDA: the covariance matrices of data classes are assumed to be different in the MD criterion while they are considered to be the same for both data classes in the LDA. Therefore, optimizing the MD in selection/extraction of the most discriminative spectral regions does not guaranty the optimal performance for the linear classifier. For instance, for 20 spectral features extracted, Approach 1A provides the largest MD (see Fig.3d) but worse performance of the LDA than for the PCA (see Fig. 3a).

For our autofluorescence spectral data, we do not observe any difference between Approaches 2A and 2B: it is not important whether the extracted spectral regions do overlap or do not (see Fig. 3b). However, we see that Approach 2A is more beneficial than Approach 1A: using less spectral wavelengths in selected spectral bands is better than when all spectral wavelengths are used in the partitioning of spectra (see Fig. 3c). In addition, we should mention that often both Approaches 1B and 2B tend to select/ extract very narrow spectral bands consisting only of 1-3 pixels/wavelengths which may introduce more noise fluctuations than a real difference between data classes. This may happen due to the overtraining that occurs when the same training set is used to construct the LDA and to optimize the MD criterion in the spectral regions extraction.

In order to avoid the shortages of the MD criterion mentioned above and the overtraining when extracting discriminative spectral regions, other criterion should be used. The alternative may be the apparent error (AE) (the classification error on the training set) of the linear classifier. This criterion is more close to the LDA than the MD criterion what concerns assumptions on the data classes distributions. The elusion of overtraining can be done by bootstrapping [6] the training set when calculating the apparent error. Namely, we bootstrap the training set B times constructing a linear classifier on each bootstrap replicate of the training set and calculate the average classification error of these B bootstrapped classifiers on the original training set.

Let us now consider the differences in the performance of the LDA caused by applying the AE criterion instead of the MD to extract informative spectral bands. In our

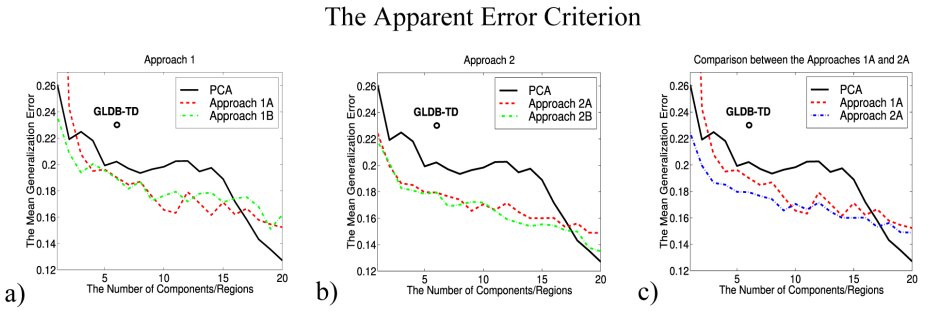

Fig. 4. The mean generalization error (GE) of LDA when the discriminative spectral regions for autofluorescence spectra are extracted/selected by using the apparent error (AE) criterion in Approach 1 (plot a) and Approach 2 (plot b). In plot (c) the comparison is made between Approaches 1A and 2A. The standard deviation of the mean GE is around 0.01. GLDB-TD denotes the performance of a Top-Down variant of the Generalized Local Discriminant Bases using an apparent error criterion.

simulations we use B=5 bootstrap replicates of the training set to calculate the apparent error. Figures 4 and 5 illustrate that using the AE criterion is indeed more beneficial than using the MD criterion: the performance of all four Approaches (1A, 1B, 2A and 2B) is improved. In figure 4 also the AE result for the GLDB is shown. In Approaches 1B and 2B, wider spectral bands on average are found to be optimal when the AE criterion is applied instead of the MD criterion for extracting discriminative spectral regions. Also, the previously made observations hold on: using a subset of spectral wavelengths in extracted spectral bands is more preferable than using them all in the optimal partition of spectra; it seems to be unnecessary to extract uncorrelated spectral bands in order to achieve the best classification results.

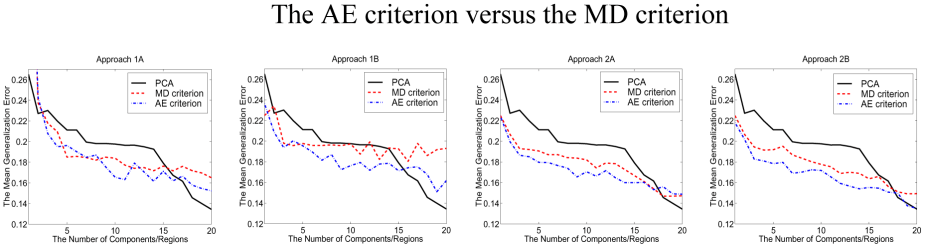

Fig. 5. The performance of LDA calculated on principal components (PCA) and on spectral regions selected/extracted by using the Mahalanobis Distance (MD) and the Apparent Error (AE) criteria for autofluorescence spectra.

5 Conclusions

Summarizing simulation results presented in the previous section, we can conclude the following. The selection/extraction of discriminative spectral bands may be more beneficial than the standard unsupervised feature selection/extraction methods (e.g.,

PCA) that do not assume the connectivity of the neighbouring wavelengths in spectral data.

In order to gain more advantage from the selection/extraction of informative spectral regions, the optimization criterion to select/extract discriminative spectral bands should be adjusted to the classification rule used to solve the problem. When selecting informative spectral regions, overtraining may be avoided if the evaluation criterion measures a performance on a different than the training dataset. For this purpose, the bootstrapped training set or an additional validation dataset can be used.

It follows from our experiments that it is advantageous to perform a multi-variate selection of wavelength groups, compared to a uni-variate approach such as Generalized Local Discriminant Bases (GLDB) utilizing the same criterion. It is also useful to exclude uninformative wavelengths from the spectral regions.

Interestingly, we have found out that extracted informative spectral bands do not need to be uncorrelated. Classifiers such as a linear discriminant assuming normal densities will find a good separation boundary even for correlated features. Feature extraction techniques like GLDB identify groups of non-overlapping wavelengths. Our finding suggests that overlapping groups of wavelengths may provide discriminative representation as well.

Acknowledgments

This work was supported by the Dutch Technology Foundation (STW), grant RRN 5316, and the Dutch Cancer Society (Nederlandse Kanker Bestrijding), grant 99-1869.

References

1. Jain, A.K., Chandrasekaran, B.: Dimensionality and Sample Size Considerations in Pattern Recognition Practice. In: Krishnaiah, P.R., Kanal, L.N. (eds.): Handbook of Statistics, Vol. 2. North-Holland, Amsterdam (1987) 835-855.
2. Nikulin, A., Dolenko, B., Bezabeh, T., Somorjai, R.: Near Optimal Region Selection for Feature Space Reduction: Novel Preprocessing Methods for Classifying MR Spectra, *NMR in Biomedicine* **11** (1998) 209-216.
3. De Veld, D.C.G., Skurichina, M., Witjes, M.J.H., et.al. Autofluorescence Characteristics of Healthy Oral Mucosa at Different Anatomical Sites, *Lasers in Surgery and Medicine*, **32** (2003) 367-376.
4. Fukunaga, K.: Introduction to Statistical Pattern Recognition. Academic Press (1990) 400-407.
5. Friedman, J.H.: Regularized Discriminant Analysis. Journal of the American Statistical Association (JASA) **84** (1989) 165-175.
6. Efron, B., Tibshirani, R.: An Introduction to the Bootstrap. Chapman and Hall, New York (1993).
7. Kumar, S., Ghosh, J. and Crawford, M.M., Best-Bases Feature Extraction Algorithms for Classification of Hyperspectral Data, IEEE Transactions on Geoscience and Remote Sensing, **39** (2001) 1368 - 1379.
8. P. Paclik, S. Verzakov, and R. P. W. Duin. Hypertools: the toolbox for spectral image analysis. Technical report, Pattern Recognition Group, TU Delft, The Netherlands, Dec. 2003.